Preprints.org

Article

Not peer-reviewed version

Aligning with Whom? A Neutrosophic Paraconsistent Framework for Value Alignment Under Irreducible Pluralism

Maikel Y. Leyva-Vázquez * , Dayron Rumbaut Rangel , Lorenzo J. Cevallos-Torres , Alexis Matheu Pérez , Florentin Smarandache

Posted Date: 30 July 2026

doi: 10.20944/preprints202607.2190.v1

Keywords:

value alignment; paraconsistent logic; neutrosophic logic; pluralistic alignment; Constitutional AI; large language models

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Aligning with Whom? A Neutrosophic Paraconsistent Framework for Value Alignment Under Irreducible Pluralism

Maikel Y. Leyva-Vázquez [1,2,*], Dayron Rumbaut Rangel [1], Lorenzo J. Cevallos-Torres [1], Alexis Matheu Pérez [3] and Florentin Smarandache [4]

[1] Universidad Bolivariana del Ecuador, Ecuador
[2] Universidad de Guayaquil, Ecuador
[3] Universidad Bernardo O'Higgins, Santiago, Chile
[4] University of New Mexico, Gallup, NM, USA
* Correspondence: mleyvaz@gmail.com

Abstract

Contemporary alignment methods—RLHF, Constitutional AI, Direct Preference Optimization—share an implicit epistemological commitment: that diverse human values can be aggregated into a single utility function or constitution fixed at training time. When principles are elicited from genuinely pluralistic populations, structurally irreducible contradictions emerge; current pipelines resolve them by consensus filters that function, in effect, as majority vote. We argue that this *alignment collapse* is not an implementation detail but a representational limit: standard uncertainty formalisms (Bayesian, fuzzy, Dempster–Shafer, imprecise probability), in their dominant normalized forms, do not preserve well-supported contradictory evidence without collapsing it into a single ordering. We propose a paraconsistent alternative in which reward signals, policy outputs, and constitutional evaluations are encoded as neutrosophic triplets $(T, I, F) \in [0,1]^3$ with no sum constraint, so that $T + F > 1$ becomes a representable—and auditable—state of genuine value conflict. We reformulate reward modeling and constitutional aggregation in this framework, introduce a deliberation-threshold decision functional, and report preliminary empirical evidence: across 2,320 evaluations of six frontier model families, models violate the classical sum constraint ($T + I + F > 1$) on 62–100% of contested baseline prompts and never on tautological controls, while the strict paraconsistent signature ($T + F > 1$) concentrates selectively on ethical-conflict items (92%) and expands under audit protocols (71%); which conflicts become visible depends strongly on which model adjudicates them (paired Cochran's Q, $p \approx 10^{-12}$). We specify a two-arm training experiment that would confirm or refute the framework's central operational claim, and close with four testable hypotheses and a research agenda that treats the question *aligning with whom?* as political and epistemic before it is technical.

Keywords: value alignment; paraconsistent logic; neutrosophic logic; pluralistic alignment; Constitutional AI; large language models

1. Introduction

Prominent alignment pipelines in production today answer a question that is rarely asked out loud: *whose* values does the system align with? Reinforcement Learning from Human Feedback trains a single reward model over aggregated annotator preferences [1]; Constitutional AI operates on one fixed constitution [2]; Direct Preference Optimization distills pairwise preferences into one implicit reward [3]; dialogue-safety systems such as Sparrow encode one rulebook [4]. Each of these methods presupposes that the diversity of human values can be compressed—by averaging, voting, or consensus filtering—into a single ordering fixed at training time.

The presupposition fails precisely where alignment matters most. When constitutional principles are elicited from genuinely pluralistic populations, structurally irreducible contradictions emerge:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

principles that majorities in one community affirm and majorities in another reject, with both positions internally coherent and well-evidenced. Anthropic's Collective Constitutional AI experiment [5], which elicited constitutional statements from roughly 1,000 U.S. participants, is the most transparent demonstration to date: contested statements were filtered through group-aware consensus before training, so that principles lacking cross-group agreement—the pool in which pluralistic information, where present, resides—never reached the model. Group-Aware Consensus is more sophisticated than simple majority vote, and was designed to protect against majority tyranny; but it remains an aggregation operator. What survives is what converges.

We call this step the *alignment collapse*: the moment at which a distribution of genuinely conflicting evaluations is projected onto a single ordering. Two observations motivate this paper. First, the collapse is an implicit epistemology, not an algorithmic necessity: it encodes the assumption that legitimate alignment requires a unique policy, and therefore that contradictions among values *must* be resolved before deployment. Second, the collapse has political consequences that the decolonial critique of AI has named with precision—the silencing of minority perspectives, indigenous epistemologies, and Global South value frameworks whose coherence does not depend on convergence with the Global North [6–8]. What that critique has articulated philosophically, however, it has not formalized mathematically. Conversely, the formal alignment community has sophisticated machinery but treats value conflict as noise to be regularized away. Recent work on pluralistic alignment [9] maps the destination—systems that answer with the distribution rather than the mode—but leaves open which formal object should carry irreducible contradiction through training and inference.

This paper proposes such an object. It is a position paper: it advances a formal proposal together with preliminary mechanism checks, with the powered experiments pre-specified but left to companion work. Our contributions are:

- a diagnosis of why standard uncertainty formalisms cannot represent well-supported value contradiction (Section 3);
- a paraconsistent alignment framework in which rewards, constitutional evaluations, and outputs are neutrosophic triplets (T, I, F) with no sum constraint, including paraconsistent reward tuples, a pluralism-preserving constitutional aggregation, and a deliberation-threshold decision functional (Section 4);
- preliminary empirical evidence that frontier LLMs already emit measurable paraconsistent signatures, selectively on value-conflict prompts—evidence that the representational gap is not hypothetical (Section 5);
- a pre-specified two-arm experiment that isolates and tests the framework's central operational claim (Section 6), plus four falsifiable hypotheses and a research agenda offered to this symposium's communities as an invitation rather than a conclusion (Section 7).

2. The Hidden Monism of Contemporary Alignment

The alignment stack, viewed end to end, performs three successive compressions. *Elicitation* compresses lived normative positions into preference labels or constitutional sentences. *Aggregation* compresses the resulting distribution into a single reward model or constitution—by majority, by consensus filter, or by curator fiat. *Optimization* compresses again, training a policy against the scalar signal. Each stage is individually reasonable; jointly they guarantee that whatever pluralism entered the pipeline, a monism exits it.

The middle compression is the least examined. In RLHF, inter-annotator disagreement is treated as label noise and averaged [1]. In Collective Constitutional AI, statements below the consensus threshold are excluded [5]. In DPO, conflicting preference pairs pull the implicit reward toward the majority direction [3]. In every case the operative assumption is that disagreement is an obstacle between the data and the model, rather than information the model should carry.

The assumption is untenable on formal grounds alone. Impossibility results in social choice and fair classification already show that natural desiderata for aggregating conflicting judgments cannot

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

be jointly satisfied [10]; there is no neutral aggregation waiting to be found. The choice of operator is a normative commitment, made silently, at the layer of the pipeline least visible to those affected by it. The question *aligning with whom?* is therefore not answered by current systems—it is dissolved by them, one consensus filter at a time.

3. Why Standard Uncertainty Frameworks Cannot Help

A natural response is that uncertainty quantification already provides the machinery: keep the disagreement as a distribution. The difficulty is that the dominant, normalized implementations of the standard formalisms are *paraconsistency-intolerant*: they do not preserve well-supported contradictory evidence as a stable, independently auditable state.

Bayesian probability enforces $P(A) + P(\neg A) = 1$; strong evidence for both A and $\neg A$ must be renormalized into moderate evidence for each, erasing the signature of conflict. **Fuzzy and intuitionistic fuzzy sets** [11] permit a hesitancy margin but retain the constraint $\mu + \nu \leq 1$: the contradictory region is defined out of the representation. **Dempster–Shafer theory** [12,13] famously misbehaves exactly where we need it: under highly conflicting evidence, Dempster's rule produces counterintuitive combinations, and the standard remedies discount or redistribute the conflict mass—again, removal rather than representation. **Imprecise probability** [14] widens the interval under conflict, conflating "the evidence is contradictory" with "the evidence is scarce"—two states that demand different downstream behavior. Non-normalized variants do exist—the Transferable Belief Model retains conflict mass explicitly [15], and belief functions have been developed over Belnap–Dunn logic—so our claim is not that no formalism can represent conflict; it is that the normalized, scalarized implementations that dominate alignment pipelines do not carry an independently auditable degree of support for both a proposition and its negation through training and inference.

Paraconsistent logics were developed precisely to reason soundly in the presence of contradiction without explosion [16–18], and Belnap–Dunn four-valued logic gave the canonical computational reading: a proposition may be *both* told-true and told-false [19]. What these traditions supply qualitatively, neutrosophic logic [20] supplies in graded, optimization-compatible form. A neutrosophic assertion about proposition A is a triplet

$$A \mapsto (T_A, I_A, F_A) \in [0,1]^3, \tag{1}$$

with *no sum constraint*. T is the degree of directly supporting evidence; F the degree of directly contradicting evidence; I the irreducible indeterminacy—the epistemic gap that neither supports nor contradicts. In normative applications, T and F are read as support for normative *admissibility* and *inadmissibility* respectively, rather than as alethic truth values; the formalism is unchanged, only its interpretation. The intuitionistic fuzzy framework [11] is embedded as the constrained subspace $T + F \leq 1$ with $I = 1 - T - F$, and classical probability as the further restriction $I = 0$, $T + F = 1$: the framework strictly generalizes both. The distinctive region is $T + F > 1$: *hyper-truth*, the graded representation of paraconsistent evidence in which both the affirmation and the negation of A are substantively supported. For value alignment this is not an exotic corner case; it is the normal condition of principles elicited across cultures.

4. A Paraconsistent Alignment Framework

4.1. Paraconsistent Reward Tuples

Classical RLHF trains a policy π_θ against a scalar reward $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. We propose replacing the scalar with a *paraconsistent reward tuple*

$$R_N(s,a) = \big(T(s,a),\ I(s,a),\ F(s,a)\big) \in [0,1]^3, \tag{2}$$

where $T(s,a)$ is the degree to which action a in state s satisfies the panel of normative evaluators, $F(s,a)$ the degree to which it violates evaluators whose validity is not in question, and $I(s,a)$ the residual indeterminacy. Policy optimization becomes a Pareto problem over $(T, -F)$ subject to an *indeterminacy*

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

budget on I. The structural difference from scalarization is that the admissible region retains states where T and F are simultaneously high: the optimizer is not permitted to pretend that a culturally contested action is merely mediocre.

4.2. Constitutional Paraconsistency

Let each constitutional principle C_i emit a neutrosophic evaluation of a candidate response r: $C_i(r) = (T_i, I_i, F_i)$. Instead of a weighted scalar sum, we aggregate with an operator that *preserves* conflict when principles genuinely disagree. Our default is the pessimistic–pluralist aggregation

$$R_C(r) = \Big(\max_i T_i,\ \operatorname{avg}_i I_i,\ \max_i F_i \Big), \tag{3}$$

which records that some principle strongly endorses r *and* some principle strongly condemns it, rather than splitting the difference. The two maxima may be attained by *different* principles, and that is the intended reading: the aggregate triplet is a panel-level summary, so $T + F > 1$ at the aggregate certifies *inter-principle* conflict—somewhere in the panel one principle strongly endorses r while another strongly condemns it—not a contradiction inside any single principle. The per-principle triplets $C_i(r)$ are retained alongside the aggregate, so the finer diagnosis (which principles clash, and whether any is internally conflicted) is never lost; what the summary alone gives up is the ability to distinguish one internally conflicted principle from two opposed ones, which is why the framework treats the disaggregated panel, not the summary triplet, as the object of record. The max is also deliberately exposed: a single misspecified principle or adversarial annotation can drive F up and block aggregate endorsement. Robust variants—community-level rather than annotator-level maxima, minimum-support thresholds before a principle enters the panel, or upper quantiles in place of the max—trade some non-erasure for robustness; because provenance is retained, choosing among them is a policy decision within the framework, not a departure from it. Alternative operators (convex Pareto, median, geometric mean) encode different ethical commitments; the choice among them is an empirical and deliberative question, not a technical default. Critically, at inference time the model is allowed to *surface* the conflict—to answer with the disagreement—instead of being forced through a single ordering.

4.3. A deliberation-Threshold Decision Functional

Preserving contradiction does not mean refusing to act. For deployment contexts that require a decision, we define for each admissible action a evaluated by communities (or principles) $i = 1, \ldots, k$ the functional

$$\sigma(a) = \min_i \big[\, T_i(a) - F_i(a) - \Omega_i \cdot I_i(a) \,\big], \tag{4}$$

where $\Omega_i \geq 0$ is community i's *indeterminacy-aversion coefficient*, which we also call its deliberation threshold: the weight community i places on unresolved indeterminacy when judging admissibility. Larger Ω_i penalizes acting under indeterminacy more heavily; $\Omega_i = 0$ means community i is indifferent to it. An action is *pluriversally admissible* when $\sigma(a) \geq 0$—no community is asked to accept an action whose net supported value, discounted by that community's own tolerance for indeterminacy, is negative. The min is a Rawlsian floor rather than a utilitarian sum: it cannot be bought off by enthusiasm elsewhere. When no action satisfies $\sigma(a) \geq 0$, the functional's output is not paralysis but a *typed* escalation: the system reports which communities' evaluations block which actions, and why—machine-readable input for exactly the deliberative processes (citizen assemblies, Pol.is-style elicitation) that this symposium examines.

4.4. A Worked Example

To make the operator concrete, consider a deployment query whose right answer is genuinely contested across communities—whether a mature minor may consent to a medical procedure without parental approval—and a candidate "yes, with safeguards" response. An autonomy-centered principle returns $C_1 = (0.80, 0.10, 0.10)$; a family-integrity principle returns $C_2 = (0.20, 0.10, 0.70)$. Pessimistic–

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

pluralist aggregation (Eq. 3) gives $R_C = (\max_i T_i, \text{avg}_i I_i, \max_i F_i) = (0.80, 0.10, 0.70)$, so $T + F = 1.50 > 1$: the aggregate is flagged as genuine value conflict rather than smoothed into a lukewarm midpoint. Applying the decision functional (Eq. 4) with thresholds $\Omega_1 = \Omega_2 = 1$ gives $\sigma_1 = 0.80 - 0.10 - 0.10 = 0.60$ but $\sigma_2 = 0.20 - 0.70 - 0.10 = -0.60$, hence $\sigma(a) = \min(0.60, -0.60) = -0.60 < 0$. The action is *not* pluriversally admissible; and because the floor is a min rather than a sum, no amount of extra enthusiasm from the autonomy community can buy off the family-integrity community's veto. Rather than silently pick a side, the system emits a typed escalation naming C_2 as the blocking evaluation, with $T + F = 1.5$ (against $I = 0.10$) as the diagnostic that the block reflects substantive conflict, not mere indeterminacy. Scalar reward erases exactly this: averaged to one number, the same case reads as a mild ≈ 0.5 and ships.

4.5. What This Buys, Concretely

Three properties distinguish the framework from distribution-preserving alternatives. *Auditability*: $T + F > 1$ is a measurable flag of genuine conflict, distinct from ignorance (I high) and from consensus ($T + F \leq 1$); regulators and affected communities can query where a system's constitution is internally contested. *Non-erasure*: minority condemnation survives aggregation by construction ($\max_i F_i$), which is the formal content of the decolonial demand that dissent not be averaged away. *Gradedness*: unlike four-valued logic, every component is continuous, so the framework is a candidate multi-head alternative to scalar reward heads—compatible with gradient-based training, though it requires a multi-objective training procedure rather than a drop-in substitution.

5. Preliminary Empirical Evidence

The framework presupposes that frontier models produce externally elicitable reports in the paraconsistent region—that $T + F > 1$ is a candidate empirical signature, not only a theoretical possibility. Three probes—all from open, reproducible pipelines—bear on this. Two distinct signatures are scored throughout. The *sum-constraint violation* $T + I + F > 1$ marks any departure from normalized, simplex-like reporting. The strict *paraconsistent signature* $T + F > 1$—hyper-truth in the sense of Section 3—additionally requires that support and counter-support jointly exceed what a normalized reporter could emit. The second implies the first; the converse fails (e.g., $(0.2, 0.9, 0.1)$ violates the sum constraint out of sheer indeterminacy, with no conflict at all).

Setup: elicitation and calibration. We elicit *unconstrained* neutrosophic judgments: each prompt presents a statement and asks the model, through a fixed instruction template, to return numeric degrees of truth (T) and falsity (F)—and, under per-attribute protocols, indeterminacy (I)—as *separate* quantities rather than a single confidence score; values are parsed from the model's structured output with no post-hoc calibration or rescaling. Calibration is instead checked externally, in two ways: tautological items serve as negative anchors, where a well-behaved reporter should return $T + F \leq 1$ (and every vendor does; see Table 1), and two prompt-ablation variants guard against wording artifacts. The multi-vendor study spans six frontier families, five epistemic phenomena (paradox, ignorance, vagueness, ethical conflict, contingency), and five elicitation protocols (a baseline, a single-vendor replication protocol, and three plithogenic overset/offset variants), with ten repetitions per cell [21].

Statistical treatment. We report signature incidence with Wilson 95% confidence intervals and correlate against the uncertainty baselines with Pearson and Spearman coefficients. For the evaluator crossover, where the *same* source responses are scored by three judges, we use paired tests—Cochran's Q across the three judges and exact McNemar tests for pairwise contrasts—on the complete-triple subset, rather than treating judges as independent groups. Two caveats are stated explicitly. First, each per-vendor cell in Table 1 pools repeated elicitations of a shared item set (five epistemic phenomena $\times$ ten repetitions), and the Wilson intervals treat those 50 evaluations as independent Bernoulli trials; since repetitions of the same prompt are positively correlated (87% of vendor–item cells are unanimous across all ten repetitions), the intervals are anti-conservative with respect to the item population and should be read as descriptive summaries of evaluation-level incidence rather than inferential

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

guarantees. Second, the effective sample is closer to the number of items than of evaluations, so we present these results as existence proofs, not effect-size estimates.

Models do not report like normalized reporters. Across 2,320 parsed evaluations [21], the sum-constraint violation $T + I + F > 1$ appeared under the baseline protocol in **62–100%** of contested evaluations *in every one of the six vendor families* (Table 1), which argues against the objection that the phenomenon is an artifact of any single lab's training pipeline. On tautological control items ($n = 90$), the rate was **0.0% across all vendors**—under both signatures—ruling out this simple format artifact in our controls. The violation rate also showed no detectable association with either mainstream uncertainty baseline—semantic entropy [22] ($r = -0.10$, 95% CI $[-0.44, 0.27]$, $p = 0.60$) and SelfCheckGPT [23] ($r = -0.12$, 95% CI $[-0.46, 0.25]$, $p = 0.53$; $n = 30$ vendor–phenomenon cells)—though the intervals are wide, so we claim an absence of detected association, not demonstrated orthogonality.

The strict paraconsistent signature is selective for value conflict. The stronger signature $T + F > 1$ is not ubiquitous—and its distribution is the point. At baseline it concentrates almost entirely on the ethical-conflict items (**92%** of those evaluations), while remaining at or near zero for logical paradox, epistemic ignorance, and future contingency ($\leq 3\%$) and at exactly zero on tautologies (0/90). Under the plithogenic overset audit protocol, it generalizes to **71%** of evaluations overall. This is the profile the framework predicts: models emit jointly strong support and counter-support precisely where genuine normative conflict lives, not as a blanket response style.

Table 1. Sum-constraint violation ($T+I+F > 1$) rates under the baseline neutrosophic protocol, by vendor, with Wilson 95% confidence intervals (descriptive; see Statistical treatment) [21]. Every frontier family departs from normalized reporting on contested prompts; none does on tautological controls, under either signature.

Vendor (baseline model)	$T+I+F>1$ rate	95% CI
Anthropic (claude-sonnet-4)	100% (50/50)	[92.9, 100]
Alibaba (qwen3-235b)	100% (50/50)	[92.9, 100]
DeepSeek (deepseek-chat)	88% (44/50)	[76.2, 94.4]
Meta (llama-4-maverick)	76% (38/50)	[62.6, 85.7]
OpenAI (gpt-4o)	66% (33/50)	[52.2, 77.6]
Mistral (mistral-medium-3.1)	62% (31/50)	[48.2, 74.1]
Tautology controls (pooled)	0% (0/90)	[0, 4.1]

Whose evaluation? The auditor effect. A crossover design—four source vendors whose responses were each re-scored by three judge models—shows that the *identity of the evaluating model* shifts which responses are classified as hyper-true (Table 2). On the 133 source responses scored by all three judges, the paired hyper-truth ($T+F > 1$) rates are 0.44 (Claude), 0.34 (Llama), and 0.14 (GPT-4o): Cochran's $Q = 55.5$, df $= 2$, $p = 9.1 \times 10^{-13}$, with every pairwise exact McNemar contrast significant (Claude vs. GPT-4o: 41+0 discordant pairs, $p = 9.1 \times 10^{-13}$; Llama vs. GPT-4o: $p = 1.4 \times 10^{-6}$; Claude vs. Llama: $p = 2.6 \times 10^{-3}$). The asymmetry is stark: GPT-4o never assigned hyper-truth to a response that Claude did not. The evaluator is not a neutral instrument; it is a participant with a standpoint. This is the alignment collapse observed *in vitro*: which contradictions become visible depends on who is asked to adjudicate them—a result any aggregation-based pipeline inherits silently, and that a paraconsistent pipeline is designed to expose.

A directional-bias null, read from the weights. One might expect the alignment-relevant signal to be a *directional* bias—models systematically privileging, say, North-elite over Global-South sources. We tested this directly, reading the internal (T, I, F) state of the residual stream with a pre-fitted Jacobian lens [24] across the Qwen3.5 family at 0.8B and 4B (a 27B replication is underway and is *not* included in the results reported here) on 480 matched-template stimuli in which a Northern and a Southern source affirm and deny the *same* true claim, with the sentence frame held byte-identical across conditions. Neither a geographic-credence discount nor a prestige-over-truth deference reached significance in the hypothesized direction at either scale, in English or Spanish (if anything, the 4B effect was weakly

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

reversed). We report this null deliberately, because it redirects the problem. The difficulty this paper formalizes is not that aligned models lean toward one pole—at these scales they do not—but that the evidence itself is paraconsistent ($T + F > 1$), and that *which* conflicts become visible depends on who adjudicates them (the auditor effect above). Bias *detection* asks the wrong first question; faithful *representation* of irreducible conflict is the prior problem, and it is the one this framework is built to solve.

Table 2. The auditor effect: paired hyper-truth ($T+F > 1$) rates assigned to the *same* source responses by three judge models (columns), restricted to responses scored by all three judges (n per source). Judge identity shifts detection threefold (Cochran's $Q = 55.5$, $p = 9.1 \times 10^{-13}$; all pairwise McNemar $p < 0.01$) [21].

Source vendor	Claude	Llama	GPT-4o	n
Alibaba	0.33	0.36	0.19	42
Anthropic	0.53	0.32	0.11	47
DeepSeek	0.62	0.42	0.12	24
Mistral	0.25	0.25	0.10	20
Pooled	0.44	0.34	0.14	133

These results are preliminary—single-language for the elicitation probes, with modest per-item samples—and we treat elicited $T + F > 1$ as a *candidate* paraconsistent signature until it is validated against culturally diverse human annotation panels (H1): high self-reported T and F can also reflect scale misunderstanding, instruction-following pressure, or prompt framing. What the controls establish is that the signature is selective and format-robust; what remains open is whether it tracks the value conflicts human communities actually recognize. Read this way, the probes ground the central claim in its honest form: where value conflict is genuine, deployed models' elicited reports already land in the region that dominant formalisms do not represent.

6. An Executable Test of Operational Gain

The evidence above shows that models occupy the paraconsistent region; it does not yet show that *training* in that representation helps. That is the framework's central operational claim (H2), and it is directly testable. We specify the experiment, and report a miniature pilot as a mechanism check; we would report a refutation as readily as a confirmation.

Design. From one open instruct model M_0 (a 7–8B base) we train policies holding data, compute, and optimizer fixed. *Arm A (baseline)* reproduces consensus-filtered Constitutional AI: constitutional statements below a group-aware-consensus threshold are dropped, and the surviving principles define a single scalar reward. *Arm B (treatment)* keeps every principle, emits a neutrosophic evaluation (T_i, I_i, F_i) per principle, aggregates with the pessimistic–pluralist operator (Eq. 3), and optimizes on the Pareto front of $(T, -F)$ under an indeterminacy budget. A vs. B is a *bundled* contrast—it changes principle retention, reward representation, and aggregation together—which mirrors deployed practice but cannot attribute a difference to any single ingredient. The confirmatory design is therefore a 2×2 factorial, {consensus-filtered, all principles} $\times$ {scalar, neutrosophic (T, I, F)} reward, whose main effects separate the contribution of principle retention from that of representation and whose interaction captures their joint effect. We draw the contested constitution and its group-level support from the public Collective Constitutional AI corpus [5], whose sub-threshold statements form a candidate pool of low-consensus, group-contested, or otherwise excluded principles, from which genuinely minority-endorsed principles must be identified before training.

Metrics. The primary outcome is the *cultural collapse rate*: on held-out prompts where the elicited communities disagree, the fraction of responses valid under a minority-endorsed principle that the policy nonetheless suppresses or penalizes relative to the majority position. The pre-specified comparison is one-sided, collapse$(B) <$ collapse(A), with bootstrap confidence intervals. Two secondary metrics guard the claim: *helpfulness parity*, assessed as equivalence rather than a failed difference test—we pre-specify a margin δ and require two one-sided tests (TOST) to bound the between-arm win-rate

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

difference within $\pm\delta$ on a neutral held-out benchmark, so any collapse reduction is not bought with capability—and *auditability*—the rate at which Arm B correctly raises a $T + F > 1$ flag on items an annotation panel marks as genuinely contested.

Isolation and falsification. Because all arms share M_0, data, and compute, differences are attributable to the manipulated factors; the two-arm contrast instantiates the diagonal corners of the factorial, which is why we read it as a deployed-practice comparison and reserve factor attribution for the full 2×2. The experiment *refutes* H2 if the neutrosophic arms fail to lower the collapse rate at helpfulness equivalence. The supporting stack—(T, I, F) reward heads, the aggregation library, and the culturally annotated evaluation suite—is under active development and will be released openly.

Pilot run (mechanism check, not a confirmatory test). We executed a miniature version of this design: Qwen2.5-0.5B-Instruct, LoRA/DPO, a hand-built constitution of 16 contested and 8 consensus items (10 contested for training, 6 held out), single seed.[1] The training signal took: the consensus-filtered arm's in-sample collapse rose from a base rate of 0.80 to 1.00—Arm A *learned* to suppress the minority-endorsed response—while Arm B's remained at 0.70. On held-out contested items the direction was consistent with H2: Arm A preferred the majority over the equally legitimate minority response on 6/6 items (collapse = 1.00; consensus-filter training *generalized* the suppression), whereas Arm B did so on 2/6 (0.33), against a base rate of 0.67, with broadly comparable helpfulness on consensus items (mean response log-probability -2.09 vs. -2.26; base -2.75). With six evaluation items, one seed, and a 0.5B model, this is an existence-level illustration that the mechanism behaves as designed—consensus filtering propagates erasure, the paraconsistent objective does not—and nothing more; the powered experiment above remains the test that counts.

7. Research Agenda: Four Hypotheses

The program is organized around four falsifiable hypotheses.

H1 (Existence). Frontier LLMs systematically produce evidence signatures with $T + F > 1$ on prompts where cultural, ethical, or epistemic frameworks genuinely conflict. *Status: preliminarily supported and selective—at baseline the signature concentrates on ethical-conflict items (Section 5); needs multilingual replication with culturally diverse annotator panels to validate the candidate signature.*

H2 (Operational gain). Neutrosophic constitutional training yields policies with measurably lower *cultural collapse*—operationalized as the erasure rate of minority-endorsed evaluations—than consensus-filtered Constitutional AI, at equivalent helpfulness. *Test: the two-arm fine-tuning experiment specified in Section 6.*

H3 (Trust asymmetry). Paraconsistent explanations ("communities disagree, here is the structure of the disagreement") increase user trust more in Global South contexts than in Global North contexts. *Test: multi-country controlled user study.*

H4 (Minimality). The triplet is the minimal sufficient structure: collapsing any component measurably loses alignment-relevant information, while adding components (meta-indeterminacy, split falsity) does not measurably add it. *Test: ablation across representational variants.*

The supporting stack—an open evaluation suite of culturally annotated contested prompts in five languages, a neutrosophic aggregation library, and (T, I, F) reward-model heads—is under development and will be released openly, building on existing repositories [21,25].

8. Position Within the Symposium and Conclusion

This symposium asks *aligning with whom?* The framework proposed here gives the question a formal home rather than an answer. Its wager is that the answer current pipelines give by default—*with whoever survives the consensus filter*—is neither politically neutral nor technically necessary. Paraconsistent, three-component representations make a different default expressible: aligning with the *structure*

[1] Pilot notebook and raw results: https://github.com/mleyvaz/pluriversal-alignment-aaai2026.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

of the disagreement itself, deferring collapse from training time to deliberation time, and returning the authority over that collapse to the communities it affects, each with its own deliberation threshold Ω_i.

We offer the framework to the symposium's three communities in a spirit of productive exposure. To formal logicians: is the neutrosophic triplet the right graded carrier for Belnap–Dunn-style information states in learning systems, or is there a better one? To empirical researchers of cross-cultural elicitation: do the hyper-truth signatures we measure track the value conflicts your instruments elicit? To Global South and decolonial scholars: does non-erasure by construction, as formalized here, answer any part of the demand for epistemic justice—or does formalization itself repeat the gesture it critiques? The working groups of this symposium are the right place to find out.

Acknowledgments: The first author thanks the Asociación Latinoamericana de Ciencias Neutrosóficas and the editorial communities of *Neutrosophic Sets and Systems* and *Neutrosophic Computing and Machine Learning* for sustained exchanges that shaped this program.

References

1. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 27730–27744.
2. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073* **2022**.
3. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct preference optimization: Your language model is secretly a reward model. In Proceedings of the Advances in Neural Information Processing Systems, 2023, Vol. 36.
4. Glaese, A.; McAleese, N.; Trebacz, M.; et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* **2022**.
5. Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T.I.; Durmus, E.; Tamkin, A.; Ganguli, D. Collective Constitutional AI: Aligning a language model with public input. In Proceedings of the Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 2024, pp. 1395–1417.
6. Mhlambi, S. From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. Technical Report 2020-009, Carr Center for Human Rights Policy, Harvard Kennedy School, 2020.
7. Birhane, A. Algorithmic injustice: A relational ethics approach. *Patterns* **2021**, *2*, 100205.
8. de Sousa Santos, B. *Epistemologies of the South: Justice against Epistemicide*; Routledge, 2014.
9. Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; et al. A roadmap to pluralistic alignment. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.
10. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* **2016**.
11. Atanassov, K.T. Intuitionistic fuzzy sets. *Fuzzy Sets and Systems* **1986**, *20*, 87–96.
12. Dempster, A.P. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics* **1967**, *38*, 325–339.
13. Shafer, G. *A Mathematical Theory of Evidence*; Princeton University Press, 1976.
14. Walley, P. *Statistical Reasoning with Imprecise Probabilities*; Chapman and Hall, 1991.
15. Smets, P.; Kennes, R. The transferable belief model. *Artificial Intelligence* **1994**, *66*, 191–234.
16. da Costa, N.C.A. On the theory of inconsistent formal systems. *Notre Dame Journal of Formal Logic* **1974**, *15*, 497–510.
17. Priest, G. The logic of paradox. *Journal of Philosophical Logic* **1979**, *8*, 219–241.
18. Carnielli, W.; Coniglio, M.E. *Paraconsistent Logic: Consistency, Contradiction and Negation*; Springer, 2016.
19. Belnap, N.D. A useful four-valued logic. *Modern Uses of Multiple-Valued Logic* **1977**, pp. 5–37.
20. Smarandache, F. *Neutrosophy: Neutrosophic Probability, Set, and Logic*; American Research Press, 1998.
21. Leyva-Vázquez, M.Y.; Smarandache, F. Breaking the chains of scalar truth: Hyper-truth signatures in frontier language models. Multi-vendor empirical study. *Preprint, github.com/mleyvaz/breaking-chains-v3-empirical* **2026**.
22. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **2024**, *630*, 625–630.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2026 doi:10.20944/preprints202607.2190.v1

23. Manakul, P.; Liusie, A.; Gales, M.J.F. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 9004–9017.
24. Anthropic Interpretability Team. Verbalizable Representations Form a Global Workspace in Language Models, 2026. Transformer Circuits; Jacobian lens reference implementation, https://github.com/anthropics/jacobian-lens.
25. Leyva-Vázquez, M.Y.; Smarandache, F. Neutrosophic interval-indeterminate numbers of the form $a + bI$: A comparative analysis with interval and affine arithmetic. *Preprint, github.com/mleyvaz/neutrosophic-affine* **2026**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.